Cross-validation in the One Sample Location Problem

by

Ronald C. Pruitt
University of Minnesota
Technical Report #510
March 1988

Key words: cross-validation, jackknife, trimmed mean.

Abstract

The use of cross-validation in choosing an estimate from among the family of trimmed means is explored for the one sample location problem. The leave one out method is not optimal for this problem, and the intuitive justification for the procedure is not compelling. An intuitively more appealing procedure is examined and shown to perform poorly by asymptotically always selecting the sample mean regardless of the underlying distribution. Finally, if a trimmed mean is chosen by minimizing the Euclidean distance to the vector of leave one out trimmed means, they only obvious optimal procedure is to minimize the jackknife estimate of variance.

1. Introduction. We consider the problem of estimating θ after observing $X_1, X_2, \ldots, X_n$ independent and identically distributed according to $F(x-\theta)$ where F is a distribution function symmetric about zero. Throughout the paper we will assume without loss of generality that $\theta = 0$, but the θ will be retained in some formulas for clarity. The particular aspect of this problem we are interested in is the use of squared error cross-validation as a means of choosing an estimator from a class of potential estimators. To be more precise, let S denote the sample $X_1, X_2, \ldots, X_n$ and let $\{\hat{\theta}(\alpha,S), \alpha \epsilon A\}$ be the set of estimators under consideration where A is some index set. Suppose that $\hat{\theta}(\cdot,\cdot)$ provides a prescription for all sample sizes, and let $\tilde{S}_i$ denote the sample of size n-1 obtained by removing X_i from S. The cross-validatory choice of α is obtained by minimizing

$$CV_n(\alpha) = n^{-1} \sum_{i=1}^{n} (X_i - \hat{\theta}(\alpha,\tilde{S}_i))^2 \tag{1.1}$$

over all possible choices of α. General accounts of cross-validation (predictive sample reuse) may be found in Stone (1974) and Geisser (1975). The technique has been used in a variety of problems, both discrete and continuous, parameteric and nonparametric. Examples include Stone (1977), Hall (1981), Eastment and Krzanowski (1982), Chow, Geman, and Wu (1983), and Bowman (1984). General theoretical results on the optimality of cross-validation are contained in Bowman, Hall, and Titterington (1984), and Li (1987). The theoretical work has centered on "hard" problems in which the best estimates converge at a rate slower than $n^{-1/2}$ such as nonparametric regression and density estimation. In this paper, we consider a particular "simpler" problem and hope to indicate some of the reasons why similar optimality theorems do not hold in this case.

The cross-validation technique given in (1.1) is based on squared error loss. Other loss functions can be used, but this paper relies heavily on manipulations involving the particular form of the loss function, and the results do not transfer easily to other loss functions. The form (1.1) is also referred to as leave one out cross-validation since the observations are removed one at a time. We present results for this form in Section 3 and give results for a modified form of leave two out cross-validation in Section 4. A class of potential estimates also needs to be specified. Two approaches seem to have been taken for data adaptive estimates in this problem. One is to try and develop a fully efficient estimate by nonparametrically estimating the score function (Beran (1974), Stone (1975), Sacks (1975)); the other chooses from a small finite number of possible estimates (Hogg (1974), Geisser (1975), Prescott (1978)). An intermediate type of approach was taken by Jaeckel (1971) who looked at the family of trimmed means and adaptively chose the trimming percentage by minimizing an ad hoc estimate of the asymptotic variance. The Jaeckel procedure is asymptotically equivalent to minimizing the jackknife estimate of variance and is discussed further in Section 5. Choosing from among the family of trimmed means is an attractive compromise: the estimates do not perform well for short or long tailed distributions, but the index set is only one dimensional. The comparison with Jaeckel's procedure also provides some insight. We only consider choosing estimates from among the family of trimmed means, but some of the general comments made apply in a wider setting. Some preporatory material on trimmed means is given in Section 2.

Now assume that X_1 has a density f, and let $Y_1 < Y_2 < \ldots < Y_n$ be the order statistics of the sample S. Let S_i denote the sample of size n-1 obtained by deleting Y_i from S. We can then rewrite (1.1) as

$$(1.2)\quad CV_n(\alpha) = n^{-1} \sum_{i=1}^{n} (X_i - \hat{\theta}(\alpha,\tilde{S}_i))^2 = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{\theta}(\alpha,S_i))^2.$$

Under the second formulation of (1.2) the intuitive justification for cross-validation does not seem as strong. As a predictor for Y_1, $\hat{\theta}(\alpha,S_1)$ does not perform well, but on the other hand $\hat{\theta}(\alpha,\tilde{S}_i)$ is a reasonable predictor for X_i in the first decomposition. The two sums appear at odds because S_i retains information about the ordering of the X_i. The conditional distribution of Y_i given S_i is not the same as the unconditional distribution of Y_i. This intuitive problem does not always occur, for instance in nonparametric regression with inference conditional on the sample from the independent variable (assumed to be continuous), there is only one observation at each value of the independent variable and no ordering problems arise since no functional form is assumed for the regression function. Attempting to fix the intuitive problems in (1.2), we have tried a cross-validation procedure leaving out observations in pairs, but instead of leaving out all possible pairs when summing the loss, we leave out only pairs of the form (Y_i,Y_{n+1-i}). This balanced cross-validation is explored in Section 4 and shown to perform very poorly by asymptotically always choosing a trimming percentage of zero. The intuitive problem in (1.2) seems to be a reflection of bad intuition rather than a bad procedure.

2. Trimmed means. In this section we provide the prescription $\hat{\theta}$ and give some results on trimmed means which are used in Section 3 and 4. Let F_n and F_n^{-1} be the usual empirical distribution function and empirical quantile function for the sample S. Let F_{ni} and F_{ni}^{-1} be the corresponding functions for S_i. Define the α-trimmed mean as

$$(2.1)\quad \hat{\theta}(\alpha,S) = \hat{\theta}(\alpha) = (1-2\alpha)^{-1}\int_{\alpha}^{1-\alpha} F_n^{-1}(y)dy$$

which reduces to $(n-2k)^{-1}\sum_{k+1}^{n-k} Y_i$ if $k = d[n\alpha]$, where $d[x]$ equals the integer part of x. The general form (2.1) is not any harder to deal with after some preliminary work. Similarly define

$$(2.2)\quad \hat{\theta}(\alpha,S_i) = (1-2\alpha)^{-1}\int_{\alpha}^{1-\alpha} \bar{F}_{ni}^{-1}(y)dy.$$

We will examine (1.2) by adding and subtracting the average of $\hat{\theta}(\alpha,S_i)$, so we will be interested in quantities such as $\hat{\theta}(\alpha,S_i) - \bar{\theta}(\alpha)$ and $\hat{\theta}(\alpha) - \bar{\theta}(\alpha)$ where

$$\bar{\theta}(\alpha) = n^{-1}\sum_{i=1}^{n}\hat{\theta}(\alpha,S_i) = (1-2\alpha)^{-1}\int_{\alpha}^{1-\alpha}\bar{F}_n^{-1}(y)dy.$$

To this end, note that

$$(2.3)\quad F_n^{-1}(y)-\bar{F}_n^{-1}(y) = \begin{cases} \frac{k}{n}(Y_k-Y_{k+1}), & \frac{k-1}{n-1} < y \le \frac{k}{n} \\ \frac{n-k}{n}(Y_{k+1}-Y_k), & \frac{k}{n} < y \le \frac{k}{n-1} \end{cases}$$

Noting that the integral of this function over $[(j-1)(n-1)^{-1}, j(n-1)^{-1}]$ is zero and letting

$$(2.4)\quad E_{1n}(\alpha) = (Y_{j+1} - Y_j) + (Y_{n-j} - Y_{n-j+1}), \qquad \frac{j-1}{n-1} < \alpha \le \frac{j}{n-1},$$

we find that

$$(2.5)\quad \hat{\theta}(\alpha)-\bar{\theta}(\alpha)= \begin{cases} (1-2\alpha)^{-1}(\alpha-\frac{k-1}{n-1})\ \frac{k}{n}\ E_{1n}(\alpha), & \frac{k-1}{n-1} < \alpha \le \frac{k}{n} \\ (1-2\alpha)^{-1}\ (\frac{k}{n-1}- \alpha)\frac{n-k}{n}\ E_{1n}(\alpha), & \frac{k}{n} < \alpha \le \frac{k}{n-1} \end{cases}$$

Similar calculations with F_{ni}^{-1} show

$$(2.6)\quad F_{ni}^{-1}(y) - \bar{F}_n^{-1}(y) = \begin{cases} \frac{j}{n}\ (Y_j-Y_{j+1}), & \frac{j-1}{n-1} < y \le \frac{j}{n-1},\ j < i \\ \frac{n-j}{n}\ (Y_{j+1}-Y_j), & \frac{j-1}{n-1} < y \le \frac{j}{n-1},\ j \ge i \end{cases}.$$

The integrals in $\hat{\theta}(\alpha,S_i)- \bar{\theta}(\alpha)$ reduce to sums with many telescoping terms which simplify. Define $j=u[(n-1)\alpha]$ where $u[\,]$ signifies next greatest integer. Then define

$$Y_i^* = \begin{cases} Y_{j+1} & i \le j \\ Y_i & j < i \le n-j \\ Y_{n-j} & n - j < i \end{cases},$$

and $\bar{Y}^* = n^{-1} \sum_{i=1}^{n} Y_i^*$ whose dependence on α is suppressed. With these definitions

$$(2.7)\qquad (1-2\alpha)(\hat{\theta}(\alpha,S_i)-\tilde{\theta}(\alpha))=\begin{cases}(n-1)^{-1}(\bar{Y}^*-Y_i^*)+(\alpha-\frac{j}{n-1})\left[\frac{j}{n}E_{1n}(\alpha)+(Y_j-Y_{j+1})\right], & i\le j\\ (n-1)^{-1}(\bar{Y}^*-Y_i^*)+(\alpha-\frac{j}{n-1})\frac{j}{n}E_{1n}(\alpha), & j<i\le n-j.\\ (n-1)^{-1}(\bar{Y}^*-Y_i^*)+(\alpha-\frac{j}{n-1})\left[\frac{j}{n}E_{1n}(\alpha)+(Y_{n-j+1}-Y_{n-j})\right], & n-j<i\end{cases}$$

The end effects in (2.7) are small, and it will be convenient later to have a simple form for a relation like (2.7), so we define $\theta_i^*(\alpha)$ by

$$(1-2\alpha)(\theta_i^*(\alpha)-\tilde{\theta}(\alpha))=\begin{cases}(\alpha-\frac{j}{n-1})\left[\frac{j}{n}E_{1n}(\alpha)+(Y_j-Y_{j+1})\right], & i\le j\\ (\alpha-\frac{j}{n-1})\frac{j}{n}E_{1n}(\alpha), & j<i\le n-j,\\ (\alpha-\frac{j}{n-1})\left[\frac{j}{n}E_{1n}(\alpha)+(Y_{n-j+1}-Y_{n-j})\right], & n-j<i\end{cases}$$

to find simplified versions of (2.5) and (2.7) as

$$(2.8)\qquad \hat{\theta}(\alpha,S_i)-\theta_i^*(\alpha)=(1-2\alpha)^{-1}(n-1)^{-1}(\bar{Y}^*-Y_i^*),$$

and

$$(2.9)\qquad \max_i|\hat{\theta}(\alpha)-\theta_i^*(\alpha)|\le(1-2\alpha)^{-1}n^{-1}3D_n(\alpha),$$

where $D_n(\alpha)=|Y_{j+1}-Y_j|+|Y_{n-j}-Y_{n-j+1}|$ for $(j-1)(n-1)^{-1}<\alpha\le j(n-1)^{-1}$.

For the case of balanced cross-validation where observations are left out in balanced pairs we need formulas analogous to (2.8) and (2.9). The results for odd and even n are quite similar although the intermediate results corresponding to (2.3) and (2.6) depend on parity. Here let S_{ii} be the sample of size n-2 obtained by deleting Y_i and Y_{n+1-i} from S. Let F_{nii} and F_{nii}^{-1} be the empirical distribution and quantile function for this sample. Let m=d[n/2] and

$$\bar{\theta}_2(\alpha) = m^{-1} \sum_{i=1}^{m} \hat{\theta}(\alpha, S_{ii}) \equiv (1-2\alpha)^{-1} \int_{\alpha}^{1-\alpha} \bar{F}_{2n}^{-1}(y)dy.$$

Again it is useful to use a close substitute for $\bar{\theta}_2(\alpha)$ to simply the expressions so we define

$$g = u[(n-2)\alpha],$$
$$E_{2n}(\alpha) = (Y_{g+1} - Y_g) + (Y_{n-g} - Y_{n-g+1}),$$

and define $\theta^*_{2i}(\alpha)$ by

$$(1-2\alpha)(\theta_{2i}(\alpha) - \bar{\theta}_2(\alpha)) = \begin{cases} \left(\frac{g}{n-2} - \alpha \right) \frac{m-g}{m} E_{2n}(\alpha) & i \le g \\ \left(\alpha - \frac{g}{n-2} \right) \frac{g}{m} E_{2n}(\alpha) & g < i \le m \end{cases}$$

Then the formulas corresponding to (2.8) and (2.9) are

$$(2.10)\quad \hat{\theta}(\alpha,S_{ii}) - \theta^*_{2i}(\alpha) = (1-2\alpha)^{-1}\,(n-2)^{-1}(2\bar{Y}^*_{2m} - Y^*_{2i} - Y^*_{2,n+1-i}),$$

and

$$(2.11)\quad \max_i |\hat{\theta}(\alpha) - \theta^*_{2i}(\alpha)| \le (1-2\alpha)^{-1}\, n^{-1}\, 2|E_{2n}(\alpha)|,$$

where

$$Y^*_{2i} = \begin{cases} Y_{g+1} & i \le g \\ Y_i & g < i \le n-g, \\ Y_{n-g} & n-g < i \end{cases}$$

and $\bar{Y}^*_{2m} = (2m)^{-1} \sum_{i=1}^{m} (Y^*_i + Y^*_{n+1-i})$. These are the needed results on trimmed means for Sections 3 and 4.

3. The leave one out method. We analyze the method of choosing α to minimize (1.2). Example 3.2 of Stone (1977) is related to this problem. In that example Stone shows that for normally distributed errors and possible estimates being the mean and the median, asymptotically cross-validation chooses the mean with probability 0.4992 and the cross-validatory estimate has efficiency 0.87 which is significantly larger than that of a random estimate equal to the mean with probability 0.4992 and the median otherwise which has efficiency 0.818. Clearly the leave one out method is not asymptotically optimal for this problem, and our goal is to provide more insight into this behavior.

Rewrite (1.2) as

$$
\begin{aligned}
CV_n(\alpha) &= n^{-1} \sum_{i=1}^{n} (Y_i - \hat{\theta}(\alpha, S_i))^2 \\
&= n^{-1} \sum_{i=1}^{n} (Y_i - \theta)^2 + n^{-1} \sum_{i=1}^{n} (\hat{\theta}(\alpha, S_i) - \theta)^2 + 2n^{-1} \sum_{i=1}^{n} (Y_i - \theta)(\theta - \hat{\theta}(\alpha, S_i)) \\
&= \sigma_n^2 + L_n(\alpha) + X_n(\alpha).
\end{aligned} \tag{3.1}
$$

The first term does not depend on α, the second is an approximation to $L(\alpha,S) = (\hat{\theta}(\alpha,S)-\theta)^2$ the quantity we would like to minimize, and the third has expectation zero due to the nature of cross-validation and the fact that X_i is unbiased for θ. From this decomposition it is hard to imagine cross-validation performing very badly for any problem as long as at least one of the quantities analogous to X_i and $\hat{\theta}(\alpha,\tilde{S}_i)$ is unbiased. In general $L_n(\alpha)$ and $X_n(\alpha)$ both converge to zero, and the properties of the procedure depend on the rates at which this convergence occurs. If we let $R_n(\alpha) = E[L_n(\alpha)]$, then Li (1987) has shown that a basic type of condition necessary for an asymptotic optimality property to hold is that $nR_n(\alpha)$ diverge uniformly in α. This type of condition is sufficient since $\mathrm{Var}(R_n(\alpha)^{-1}X_n(\alpha)) = C_1 n^{-1} R_n(\alpha)^{-1}$ so if $R_n(\alpha)$ converges to zero slowly enough, $X_n(\alpha)$ converges to zero faster than $L_n(\alpha)$. In this problem, as long as the tails of X_1 are not overly long or short, $nR_n(\alpha) \to C_2$ as $n \to \infty$ and $\mathrm{Var}\,(R_n(\alpha)^{-1}X_n(\alpha)) \to C_3$. In other words, $CV_n(\alpha)$ properly normalized so that $L_n(\alpha)$ does not converge to 0 or ∞, behaves like a constant not depending on α plus a nondegenerate random variable. This is indeed the case as we shall show.

To get at the asymptotic behavior of $CV_n(\alpha)$ we make an expansion like (3.1) around $\theta_i^*(\alpha)$,

$$nCV_n(\alpha) = \sum_{i=1}^{n} (Y_i - \theta_i^*(\alpha))^2 + \sum_{i=1}^{n} (\hat{\theta}(\alpha,S_i) - \theta_i^*(\alpha))^2$$

$$(3.2) \qquad + 2 \sum_{i=1}^{n} (Y_i - \theta_i^*(\alpha))(\theta_i^*(\alpha) - \hat{\theta}(\alpha,S_i))$$

$$= \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (2Y_i - \theta_i^*(\alpha) - \bar{Y})(\bar{Y} - \theta_i^*(\alpha)) + (1-2\alpha)^{-2}(n-1)^{-2} \sum_{i=1}^{n} (Y_i^* - \bar{Y}^*)^2$$

$$+ 2(1-2\alpha)^{-1}(n-1)^{-1} \sum_{i=1}^{n} (Y_i - \theta_i^*(\alpha))\,(Y_i^* - \bar{Y}^*)$$

by using (2.8). The limits of these terms are fairly easy to obtain and are given in Theorem 3.1. We first give some necessary notation.

Let $A_\alpha(F) = \{\omega: |F(X(\omega)) - 1/2| < 1/2 - \alpha\}$ and

$$T_\alpha(F) = \int_{A_\alpha(F)} (x-\theta)^2 \, dF(x) + (F^{-1}(1-\alpha) - \theta) \int_{A_\alpha(F)^c} |x-\theta| \, dF(x).$$

Uniform convergence over the entire range of α is difficult to obtain. For any fixed range $0 < \alpha_0 \le \alpha \le \alpha_1 < 1/2$, the necessary uniform convergence is possible, but it is also possible to let $\alpha_0 \to 0$ as $n \to \infty$ in the following way. Let

$$(3.3) \quad B_n(F) = \left\{ \alpha: n^{-1} < \alpha \le \alpha_1 < \frac{1}{2},\ (F^{-1}(\alpha))^2 n^{-1/2} < K_1(n),\ n^{-1/2} f(F^{-1}(\alpha))^{-1} < K_2(n) \right\}$$

for some fixed sequences of positive constants $K_1(n)$ and $K_2(n)$ approaching zero as $n \to \infty$. If the $K_j(n)$ are chosen to approach zero more slowly than $n^{-1/2}$ then $B_n(F)$ increases to $(0,\alpha_1]$ as $n \to \infty$. Except for very short tailed distributions F, for which trimmed means are not good estimates, the first restriction in $B_n(F)$ is not the strictest for small values of α. For distributions lacking

second moment the K_1 condition will be strictest and otherwise the K_2 condition will be the most strict. For example, if F is the Laplace distribution the third condition reduces to $K_2(n)^{-1}n^{-1/2} < \alpha$. Smaller values of α are allowed for shorter tailed distributions. Finally, let

$$Y_n(\alpha) = nCV_n(\alpha) - \sum_{i=1}^{n}(X_i-\bar{X})^2 - n(\bar{X}-\hat{\theta}(\alpha))^2 - 2(1-2\alpha)^{-1}T_\alpha(F).$$

Theorem 3.1 Under the following assumptions,

A1) Let $X_1, X_2, \ldots$ be independent identically distributed random variables with a continuous distribution function F which is twice differentiable on (a,b) where

$$a = \sup\{x: F(x) = 0\}, \quad b = \inf\{x: F(x) = 1\},$$

and $F'(x) = f(x) > 0$ on (a,b). Also assume

$$\sup_{a<x<b} F(x)(1-F(x))\frac{|f'(x)|}{f^2(x)} \le \gamma \tag{3.4}$$

for some positive γ,

A2) F is symmetric around θ,

A3) $$\sup_{\alpha\epsilon B_n(F)} |F^{-1}(\alpha)|\left|\int_{-\infty}^{\infty}|x|dF_n(x) - E|X|\right| \xrightarrow{p} 0,$$

then

$$\sup_{\alpha\epsilon B_n(F)} |Y_n(\alpha)| \xrightarrow{p} 0. \qquad \square$$

The proof is given in Section 6. Assumption 1 is the assumptions of Lemma 1.4.1 of Csörgő (1983) which allows us to prove the uniform convergence of the quantile process on $B_n(F)$. Examples of distributions satisfying (3.4) are given in Parzen (1979, §9). Assumption 3 implies X_1 has first absolute moment, and in fact if F is of the form $F(x) = (-x)^{-p}$ in the left tail for $p > 1$, it suffices that $(p-1)/p > 1/4$ for A3) to hold using the second condition of (3.3). It can be shown that $Y_n(0) \xrightarrow{p} 0$ if a second moment is assumed for X_1 so the theorem can be extended to $B_n(F) \cup \{0\}$ if desired.

Asymptotically, up to constants not involving α, $nCV_n(\alpha)$ behaves as a deterministic term, $2(1-2\alpha)^{-1}T_\alpha(F)$, plus a stochastic term $n(\bar{X}-\hat{\theta}(\alpha))^2$. In the following discussion we consider only the case when X_1 has second moment since otherwise the deterministic term is poorly behaved near $\alpha=0$ and the stochastic term blows up; the squared error loss function is also not particularly appropriate for such distributions. Let $R(\alpha) = \lim_{n\to\infty} nR_n(\alpha)$, then $R(0) = 2T_0(F)$ and $2(1-2\alpha)^{-1}T_\alpha(F) < R(\alpha)$ for $\alpha > 0$ since the stochastic term has strictly positive expectation. The stochastistic term is asymptotically Chi-squared, so that the minimizer of $nCV_n(\alpha)$ behaves as the minimizer of a Chi-squared process. The cross-validatory estimate will tend to have properties reflecting both the optimal estimate and the sample mean. Cross-validation tends to choose α's which for a particular sample happen to be close to $\bar{X}$. If $\alpha = 0$ minimizes $R(\alpha)$ (Stone (1977), example 3.2), cross-validation will often choose non-zero α's but the estimate is constrained to have relatively good performance since $\hat{\theta}(\hat{\alpha})$ must be fairly close to $\bar{X}$. However if a non-zero α minimizes $(R(\alpha)$, $\hat{\alpha}$ will still vary and will still emphasize these $\hat{\theta}(\hat{\alpha})$'s which are close to $\bar{X}$. This provides some intuition for the second result in Stone's example 3.2. He there considers modulus loss with normal observations and finds

that the efficiency of the cross-validatory estimate of choice between the mean and median is lower than that of a random procedure choosing the mean and median with the same probabilities. If modulus loss cross-validation chooses α's with $\hat{\theta}(\alpha)$ close to the sample median as squared error does for the sample mean, then the cross-validatory estimate will choose the mean when it is close to the median and otherwise choose the median. This would not give good results with normal observations.

It is tempting to explain the special role of the sample mean in this development as a result of the fact that $\bar{X}$ minimizes $\sum_{i=1}^{n}(X_i-c)^2$ over c or $\sum_{i=1}^{n}(X_i-\hat{\theta}(\alpha))^2$ over α regardless of the distribution of X_1. Undoubtedly this does play a central role in the type of estimate found by cross-validation, but it does not seem to be central to the non-optimality of the procedure. The non-optimality seem to be due to the fact that the estimates converge too quickly, see also Li (1987) for some general comments on this subject.

4. Balanced cross-validation. In this section we explore a variant of cross-validation inspired by the second sum in (1.2). When trying to predict Y_1, $\hat{\theta}(\alpha,S_1)$ is a poor prediction since Y_1 has a negative bias and $\hat{\theta}(\alpha,S_1)$ has a positive bias. If this argument is compelling, a possible solution would be to leave out observations in pairs, but only balanced pairs: that is, leave out the k^{th} smallest and k^{th} largest observations for k=1,2,...,d[n/2]. This type of cross-validation will be referred to as balanced. Under balanced cross-validation, α is chosen by minimizing

$$BCV_n(\alpha) = 4^{-1}m^{-1} \sum_{i=1}^{m} (Y_i + Y_{n+1-i} - 2\hat{\theta}(\alpha, S_{ii}))^2.$$

Following (3.2) and using (2.10),

$$mBCV_n(\alpha) = 4^{-1} \sum_{i=1}^{m} (Y_i + Y_{n+1-i} - 2\bar{Y}_m)^2 + \sum_{i=1}^{m} (Y_i + Y_{n+1-i} - \theta^*_{2i}(\alpha) - \bar{Y}_m)(\bar{Y}_m - \theta^*_{2i}(\alpha))$$

$$+ (1-2\alpha)^{-2}(n-2)^{-2} \sum_{i=1}^{m} (Y^*_{2i} + Y^*_{2,n+1-i} - 2\bar{Y}^*_{2m})^2$$

$$+ (1-2\alpha)^{-1}(n-2)^{-1} \sum_{i=1}^{m} (Y_i + Y_{n+1-i} - 2\bar{Y}_m)(Y^*_{2i} + Y^*_{2,n+1-i} - 2\bar{Y}^*_{2m}),$$

where $\bar{Y}_m = (2m)^{-1} \sum_{i=1}^{m} (Y_i + Y_{n+1-i})$. The first three terms can be handled as in Theorem 3.1, but the last term is somewhat different. The last term can be shown to converge to zero and we get the following theorem.

Theorem 4.1 Under A1) - A3) of Theorem 3.1, if

$$Z_n(\alpha) = mBCV_n(\alpha) - 4^{-1} \sum_{i=1}^{n} (Y_i + Y_{n+1-i} - 2\bar{Y}_m)^2 - m(\bar{X} - \hat{\theta}(\alpha))^2,$$

then $\sup_{\alpha \in B_n(F)} |Z_n(\alpha)| \xrightarrow{p} 0.$ □

Unlike in Theorem 3.1, here we have not been able to show $Z_n(0) \to 0$ except under rather stringent conditions (X_1 bounded). This convergence will occur

when $n^{-1} \sum_{i=1}^{n} Y_i Y_{n+1-i}$ converges to $-\mathrm{Var}(X_1)$ which we surmise to hold under much weaker conditions than X_1 bounded.

The qualitative result of Theorem 4.1 and not the technical details are what is of interest. Asymptotically, the minimizer of $BCV_n(\alpha)$ is behaving as the minimizer of a Chi-squared process. The minimizer of the Chi-squared

process over the entire interval $[0,\alpha_1]$ is $\alpha = 0$ regardless of the distribution F. This intuitively appealing modification of cross-validation has produced a terrible procedure which does not even possess the weak expectated loss estimating property discussed following (3.1). Rewrite (3.1) for this procedure as

$$\begin{aligned} BCV_n(\alpha) &= 4^{-1}m^{-1}\sum_{i=1}^{n}(Y_i+Y_{n+1-i}-2\theta)^2+m^{-1}\sum_{i=1}^{m}(\hat{\theta}(\alpha,S_{ii})-\theta)^2 \\ &\qquad + 2^{-1}m^{-1}\sum_{i=1}^{m}(Y_i+Y_{n+1-i}-2\theta)(\theta-\hat{\theta}(\alpha-S_{ii})) \\ &= \gamma_n^2 + L_{2n}(\alpha) + X_{2n}(\alpha). \end{aligned} \tag{4.1}$$

The first term does not depend on α, the second still approximates $L(\alpha,S)$, but the third does not have expectation zero. The fact that $E[X_n(\alpha)] = 0$ in (3.1) depends heavily on the sum being taken over all subsets of indices of the size being removed, as well as on X_i or $\hat{\theta}(\alpha,\cdot)$ being unbiased. Without both of these conditions being satisfied the cross-validation procedure will not in general even have the weak expectation unbiased property given by $E[X_n(\alpha)] = 0$. The expansion (4.1) provides an explanation as to why balanced cross-validation seems an intuitive improvement: in balanced cross-validation we are coming closer to approximating the loss since γ_n^2, which is nearly $2^{-1}Z_n(0)$, will presumably be small, although the approximation is not very good. Section 3 shows it is better to get a good approximation of the loss plus a constant independent of α rather than attempting to approximate the loss itself. This section shows that ad hoc adjustments to cross-validation may have disastrous consequences.

5. The jackknife. Suppose we wish to find a function $f(X_i,\alpha,\tilde{S}_i)$ such that choosing α by minimizing

$$(5.1) \qquad C(\alpha) = \sum_{i=1}^{n} (f(X_i,\alpha,\tilde{S}_i) - \hat{\theta}(\alpha,\tilde{S}_i))^2$$

possesses some optimality property for this problem. The discussion will be informal, and assumptions will be made as needed. Consider the case where F has second moment and let $\sigma^2(\alpha)$ be the asymptotic variance of $\hat{\theta}(\alpha)$. Assume $nL_n(\alpha) \to \sigma^2(\alpha)$ as $n \to \infty$. In particular, we consider under what conditions

$$(5.2) \quad K_1(n)\, C(\alpha) + K_2(n) \to \sigma^2(\alpha),$$

where $K_1(n)$ and $K_2(n)$ do not depend on α. If this convergence holds and is uniform in α, then the minimization in (5.1) will possess an asymptotic optimality property under some regularity conditions.

To explore $C(\alpha)$ denote $f(Y_i,\alpha,S_i)$ by f_i, and let $\bar{f} = n^{-1}\sum_{i=1}^{n} f_i$. Then

$$C(\alpha) = \sum_{i=1}^{n} (f_i-\bar{f})^2+(1-2\alpha)^{-2}(n-1)^{-2}\sum_{i=1}^{n} (Y_i^*-\bar{Y}^*)^2 + n(\bar{\theta}^*(\alpha)-\bar{f})^2$$

$$+ \sum_{i=1}^{n} (\theta_i^*(\alpha)-\bar{\theta}^*(\alpha))(\theta_i^*(\alpha)+\bar{\theta}_i^*(\alpha)-2f_i)+2(1-2\alpha)^{-1}(n-1)^{-1}\sum_{i=1}^{n} (f_i-\theta)(Y_i^*-\theta)$$

$$+ 2(1-2\alpha)^{-1}(n-1)^{-1}n(\bar{f}-\theta)(\theta-\bar{Y}^*)$$

where $\bar{\theta}^*(\alpha) = n^{-1}\sum_{i=1}^{n} \theta_i^*(\alpha)$, and we have assumed $E[f(X_i,\alpha,\tilde{S}_i)] = \theta$. The case of biased f amounts to minimizing $n(\bar{\theta}^*(\alpha)-\bar{f})^2$ which does not work well. The second, fourth, and sixth, terms in (5.3) are $O_p(n^{-1})$ and can be ignored unless $\bar{f} = \theta_i^*(\alpha) + O_p(n^{-1})$.

First consider the case that $f(X_i,\alpha,\tilde{S}_i)$ depends only on X_i, this is a strict view of cross-validation. In this case we can make a decomposition like (3.1) to find $E[C(\alpha)] = \sigma^2(\alpha) - K_2(n)$ where $K_2(n)$ does not depend on α. The behavior is very similar to that in Section 3 since the first term in (5.3) does not depend on α, the third term will usually converge in distribution to a nondegenerate random variable, and the fifth term will converge to a constant depending on α. If we have enough structure so that the above convergences hold, then there is no hope of obtaining (5.2). Allowing $f(X_i,\alpha,\tilde{S}_i)$ to depend on α as well, but not $\tilde{S}_i$, also seems a reasonable view of cross-validation. When only one observation is left out at a time, allowing $f(X_i,\alpha,\tilde{S}_i)$ to depend on α does not seem to provide any advantage, but this is not the case when more than one observation is deleted. Write

$$(5.4)\quad CV_n(\alpha) = n^{-1}\sum_{i=1}^{n}(Y_i-\hat{\theta}(\alpha,S_i))^2 = n^{-1}\sum_{i=1}^{n}(\hat{\theta}(\alpha,Y_i)-\hat{\theta}(\alpha,S_i))^2,$$

then by deleting more observations at a time it will be possible to estimate $\sigma^2(\alpha)$ using the sum suggested by the final form in (5.4). These ideas will be fully explored in a further paper.

Finally, we give two obvious ways in which (5.2) can hold. First we note that (5.1) looks like a jackknife estimate of variance if we take $f_i = \bar{f} - \bar{\theta}(\alpha) = \theta_i^*(\alpha) + O_p(n^{-1})$. In this case $nC(\alpha) \to \sigma^2(\alpha)$ since α is bounded away from 1/2. In fact

$$(5.5)\quad nC(\alpha) = (1-2\alpha)^{-2}\, n^{-1}\sum_{i=1}^{n}(Y_i^* - \bar{Y}^*)^2 + o_p(1)$$

which suggests another obvious method to get (5.2), by taking $f(Y_i,\alpha,S_i) = Y_i^*$.

In both of these cases a multiple of $C(\alpha)$ is roughly the jackknife estimate of variance. These are the obvious solutions since it is not clear how to handle the term $n(\bar{\theta}^*)-\bar{f})^2$ unless it is zero. The sum in (5.5) is the same type of quantity as is minimized in Jaeckel (1971) for this problem. Jaeckel's $s^2(\alpha)$ is ad hoc, but all of his proofs have straightforward modifications to the case of minimizing the jackknife estimate of variance.

6. Proofs. Proof of Theorem 3.1. From (3.2),

$$|Y_n(\alpha)| \le (1-2\alpha)^{-2}(n-1)^{-2}\sum_{i=1}^{n}(Y_i^*-\bar{Y}^*)^2 + \Big|\sum_{i=1}^{n}(\hat{\theta}(\alpha)-\theta_i^*(\alpha))(2Y_i-\theta_i^*(\alpha)-\hat{\theta}(\alpha))\Big|$$

$$+2(1-2\alpha)^{-1}\Big|(n-1)^{-1}\sum_{i=1}^{n}(Y_i-\theta_i^*(\alpha))(Y_i^*-\bar{Y}^*) - T_\alpha(F)\Big|$$

$$\le (1-2\alpha_1)^{-2}(n-1)^{-2}\sum_{i=1}^{n}(Y_i^*-\theta)^2$$

$$+(1-2\alpha_1)^{-1}6D_n(\alpha)\left\{n^{-1}\sum_{i=1}^{n}|Y_i-\theta| + |\hat{\theta}(\alpha)-\theta| + \max_i|\hat{\theta}(\alpha)-\theta_i^*(\alpha)|\right\}$$

$$+2(1-2\alpha_1)^{-1}\left\{\Big|(n-1)^{-1}\sum_{i=1}^{n}(Y_i-\theta)(Y_i^*-\theta)-T_\alpha(F)\Big| + (n-1)^{-1}n|(\bar{X}-\theta)(\bar{Y}^*-\theta)|\right\},$$

using (2.8) and (2.9). It now clearly suffices to show

$$(6.1)\quad \sup_{\alpha\in B_n(F)} D_n(\alpha) \xrightarrow{p} 0,$$

$$(6.2)\quad \sup_{\alpha\in B_n(F)} |\bar{Y}^* - \theta| \xrightarrow{p} 0,$$

$$(6.3)\quad \sup_{\alpha\in B_n(F)} |\hat{\theta}(\alpha) - \theta| \xrightarrow{p} 0,$$

$$(6.4)\quad \sup_{\alpha\in B_n(F)} \left| n^{-1} \sum_{i=1}^{n} (Y_i-\theta)(Y_i^*-\theta) - T_\alpha(F) \right| \xrightarrow{p} 0,$$

and

$$(6.5)\quad \sup_{\alpha\in B_n(F)} n^{-2} \sum_{i=1}^{n} (Y_i^*-\theta)^2 \xrightarrow{p} 0.$$

In place of (6.1), (6.2), and (6.3), we could prove $D_n(\alpha)|F^{-1}(\alpha)|^{-1} \to 0$, $|\bar{Y}^*-\theta|\ |F^{-1}(\alpha)|^{-1} \to 0$, and $|\hat{\theta}(\alpha)-\theta||F^{-1}(\alpha)| \to 0$ which hold under slightly weaker conditions, but (6.1)-(6.3) are easily proved under the assumptions necessary for (6.4).

To prove (6.1)-(6.5), we require two lemmas about empirical and quantile processes.

Lemma 6.1 (Dvoretsky, Kiefer, and Wolfowitz (1956)). Let $X_1, X_2, \ldots$ be independent identically distributed with distribution function F on $\mathbb{R}$. Let F_n be the empirical distribution function of $X_1, X_2, \ldots, X_n$. Then

$$\sup_{0\le y\le 1} |F(F_n^{-1}(y)) - y| = O_p(n^{-1/2}). \qquad \square$$

This supremum is attained and has the same value as

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|.$$

Lemma 6.2 (Theorem 1.4.3 of Csörgő (1983) with $x=n^{-1}$, $x \to \infty$, and $\lambda \to \infty$). If A1) of Theorem 3.1 holds, then

$$\sup_{n^{-1} \le y \le 1-n^{-1}} f(F^{-1}(y))|F_n^{-1}(y)-F^{-1}(y)| = O_p(n^{-1/2}). \qquad \square$$

We are now ready to prove (6.1)-(6.5). For (6.1), let $j=u[(n-1)\alpha]$, $\alpha^* = j/n$, and $\alpha^{**} = (j+1)/n$. Note by symmetry it suffices to only consider $|Y_{j+1} - Y_j|$. Now

$$\sup_{B_n(F)} |Y_{j+1}-Y_j| \le \sup_{B_n(F)} |F_n^{-1}(\alpha^{**})-F^{-1}(\alpha^{**})| + \sup_{B_n(F)} |F_n^{-1}(\alpha^*)-F^{-1}(\alpha^*)|$$

$$+ \sup_{B_n(F)} |F^{-1}(\alpha^{**}) - F^{-1}(\alpha^*)|$$

Note $j/n \notin B_n(F)$, but each of these terms converge to zero using Lemma 6.2, the Mean Value Theorem, the definition of $B_n(F)$, and the continuity of f. This shows (6.1).

Next look at (6.2) and (6.3). Let $j = u[(n-1)\alpha]$ as before and Let $\tilde{\alpha} = j/n-1$ and $\underset{\sim}{\alpha} = (j+1)/(n-1)$. Then

$$|\tilde{Y}^*-\theta| \le \left| \int_{F_n^{-1}(\tilde{\alpha})}^{F_n^{-1}(1-\tilde{\alpha})} x dF_n(x) \right| + \tilde{\alpha}|F_n^{-1}(\underset{\sim}{\alpha}) + F_n^{-1}(1-\underset{\sim}{\alpha})|$$

$$= g(\tilde{\alpha}) + \tilde{\alpha}h(\underset{\sim}{\alpha}).$$

Note that $|\hat{\theta}(\alpha) - \theta| \le (1-2\alpha_1)^{-1} g(\alpha)$ so that it suffices to show

$$(6.6) \quad \sup_{B_n(F)} g(\alpha) \xrightarrow{p} 0, \text{ and } \sup_{B_n(F)} h(\alpha) \xrightarrow{p} 0,$$

to prove (6.2) and (6.3). The second statement is clear from the symmetry of F, (3.3), and Lemma 6.2. For the first statement in (6.6),

$$g(\alpha) \le \left| \int_{F^{-1}(1-\alpha)}^{F_n^{-1}(1-\alpha)} x dF(x) \right| + \left| \int_{F^{-1}(\alpha)}^{F_n^{-1}(\alpha)} x dF(x) \right| + \left| \int_{F_n^{-1}(\alpha)}^{F_n^{-1}(1-\alpha)} x(dF_n(x) - dF(x)) \right|$$

so that

$$(6.7) \quad \sup_{B_n(F)} g(\alpha) \le 2 \sup_{B_n(F)} \Big\{ \max \left[|F^{-1}(\alpha)|, |F_n^{-1}(\alpha)|, |F_n^{-1}(1-\alpha)| \right] \times \left[|F(F_n^{-1}(\alpha)) - \alpha| + |F(F_n^{-1}(1-\alpha)) - (1-\alpha)| \right] \Big\}$$

which converges to zero using (3.3), and Lemmas 6.1 and 6.2. Therefore (6.2) and (6.3) hold.

For (6.4), recall $\alpha^* = j/n$ and $\alpha^{**} = (j+1)/n$. Then the argument of (6.4) may be written

$$\sup_{B_n(F)} \left| \int_{F_n^{-1}(\alpha^*)}^{F_n^{-1}(1-\alpha^*)} x^2 dF_n(x) + F_n^{-1}(\alpha^{**}) \int_{-\infty}^{F_n^{-1}(\alpha^*)} x dF_n(x) + F_n^{-1}(1-\alpha^{**}) \int_{F_n^{-1}(1-\alpha^*)}^{\infty} x dF_n(x) \right.$$
$$\left. - \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x^2 dF(x) - F^{-1}(\alpha) \int_{-\infty}^{F^{-1}(\alpha)} x dF(x) - F^{-1}(1-\alpha) \int_{F^{-1}(1-\alpha)}^{\infty} x dF(x) \right|$$

$$\leq \sup_{B_n(F)} \left| \int_{F^{-1}(1-\alpha)}^{F_n^{-1}(1-\alpha^*)} x^2 dF(x) \right| + \sup_{B_n(F)} \left| \int_{F^{-1}(\alpha)}^{F_n^{-1}(\alpha^*)} x^2 dF(x) \right| + \sup_{B_n(F)} \left| \int_{F_n^{-1}(\alpha^*)}^{F_n^{-1}(1-\alpha^*)} x^2 (dF_n(x) - dF(x)) \right|$$

$$(6.8) \quad + \sup_{B_n(F)} \left| F_n^{-1}(1-\alpha^{**}) + F_n^{-1}(\alpha^{**}) \right| \left| \int_0^\infty x dF_n(x) \right| + \sup_{B_n(F)} \left| F_n^{-1}(\alpha^{**}) - F^{-1}(\alpha) \right| \int_{-\infty}^{\infty} |x| dF_n(x)$$

$$+ \sup_{B_n(F)} |F^{-1}(\alpha)| \left| \int_{-\infty}^{\infty} |x| dF_n(x) - E|X| \right| + \sup_{B_n(F)} \left| F_n^{-1}(\alpha^{**}) \int_{F_n^{-1}(\alpha^*)}^{F_n^{-1}(1-\alpha^*)} |x| dF_n(x) - F^{-1}(\alpha) \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} |x| dF(x) \right|.$$

We examine this decompistion term by term. The first three terms may be handled analogously to $g(\alpha)$ at (6.7). For the fourth and fifth terms the integrals do not depend on α and converge almost surely as n increases. Hence the fourth and fifth terms converge to zero using Lemma 6.2 after noting $|F^{-1}(\alpha^{**}) - F^{-1}(\alpha)|$ converges uniformly to zero on $B_n(F)$. The sixth term goes to zero by A3). The last term may be handled in the same fashion as $g(\alpha)$. Hence (6.4) converges to zero.

Using the same techniques as for (6.4),

$$\sup_{B_n(F)} \left| n^{-1} \sum_{i=1}^{n} (Y_i^* - \theta)^2 - W_\alpha(F) \right| \xrightarrow{p} 0$$

where $W_\alpha(F) = E[(X-\theta)^2 | A_\alpha(F)] + (F^{-1}(\alpha))^2 P(A_\alpha(F)^c)$. Finally note $n^{-1} W_\alpha(F) \leq 2n^{-1}(F^{-1}(\alpha))^2$ which converges to zero uniformly on $B_n(F)$. Hence (6.5) holds and the proof of Theorem 3.1 is complete. □

Proof of Theorem 4.1. The proof is similar to that of Theorem 3.1. Write

$$|Z_n(\alpha)| \le (1-2\alpha)^{-2}(n-2)^{-2} \sum_{i=1}^{m} (Y^*_{2i} + Y^*_{2,n+1-i} - 2\bar{Y}^*_{2m})^2 +$$

$$+ \Big| \sum_{i=1}^{m} (\hat{\theta}(\alpha) - \theta^*_{2i}(\alpha))(Y_i + Y_{n+1-i} - \theta^*_{2i}(\alpha) - \hat{\theta}(\alpha)\Big| + \big|m(\bar{Y} - \bar{Y}_m)(2\hat{\theta} - \bar{Y} - \bar{Y}_m)\big|$$

$$+ (1-2\alpha)^{-1}2^{-1}\Big|(n-2)^{-1}\sum_{i=1}^{m}(Y_i + Y_{n+1-i} - 2\bar{Y}_m)(Y^*_{2i} + Y^*_{2,n+1-i} - 2Y^*_{2m})\Big|$$

$$\le (1-2\alpha_1)^{-2}(n-2)^{-2}\, 2 \sum_{i=1}^{n} (Y^*_{2i} - \theta)^2$$

$$+ (1-2\alpha_1)^{-1}2E_{2n}(\alpha)\left\{ n^{-1}\sum_{i=1}^{n}|Y_i - \theta| + |\hat{\theta}(\alpha) - \theta| + \max_i |\hat{\theta}(\alpha) - \theta^*_{2i}(\alpha)| \right\}$$

$$+ \big|(Y_{m+1} - \bar{Y})(2\hat{\theta} - \bar{Y} - \bar{Y}_m)\big| + (1-2\alpha_1)^{-1}(n-2)^{-1}n\big|(\bar{X} - \theta)(\bar{Y}^*_{2m} - \theta)\big|$$

$$+ (1-2\alpha_1)^{-1}\Big|(n-2)^{-1}\sum_{i=1}^{n}(Y_i - \theta)(Y^*_{2i} - \theta) - T_\alpha(F)\Big|$$

$$+ (1-2\alpha_1)^{-1}\Big|(n-2)^{-1}\sum_{i=1}^{n}(Y_i - \theta)(Y^*_{2,n+1-i} - \theta) + T_\alpha(F)\Big|.$$

The first five terms can be handled as in Theorem 3.1, so to prove Theorem 4.1 it suffices to show

$$(6.9) \qquad \sup_{\alpha \in B_n(F)} \Big|n^{-1}\sum_{i=1}^{n}(Y_i - \theta)(Y^*_{2,n+1-i} - \theta) + T_\alpha(F)\Big| \xrightarrow{p} 0.$$

Let $g = u[(n-2)\alpha]$, $\alpha_* = g/n$, and $\alpha_{**} = (g+1)/n$. Then the argument in (6.9) may be written

$$\sup_{B_n(F)} \Bigg| \int_{\alpha_*}^{1-\alpha_*} F_n^{-1}(y)F_n^{-1}(1-y)dy + F_n^{-1}(1-\alpha_{**}) \int_{-\infty}^{F_n^{-1}(\alpha_*)} xdF_n(x) + F_n^{-1}(\alpha_{**}) \int_{F_n^{-1}(1-\alpha_*)}^{\infty} xdF_n(x)$$

$$- \int_{\alpha}^{1-\alpha} F^{-1}(y)F^{-1}(1-y)dy - F^{-1}(1-\alpha) \int_{-\infty}^{F^{-1}(\alpha)} xdF(x) - F^{-1}(\alpha) \int_{F^{-1}(1-\alpha)}^{\infty} xdF(x) \Bigg|.$$

Comparing with (6.8) we need only show

$$(6.10) \quad \sup_{B_n(F)} \left| \int_{\alpha_*}^{1-\alpha_*} F_n^{-1}(y)F_n^{-1}(1-y)dy - \int_{\alpha}^{1-\alpha} F^{-1}(y)F^{-1}(1-y)dy \right| \xrightarrow{p} 0.$$

This is easily seen since the argument in (6.10) is smaller than

$$\sup_{B_n(F)} \left| \int_{1-\alpha}^{1-\alpha_*} F^{-1}(y)F^{-1}(1-y)dy \right| + \sup_{B_n(F)} \left| \int_{\alpha}^{\alpha_*} F^{-1}(y)F^{-1}(1-y)dy \right|$$

$$+ \sup_{B_n(F)} \left| \int_{\alpha_*}^{1-\alpha_*} (F_n^{-1}(y)F_n^{-1}(1-y) - F^{-1}(y)F^{-1}(1-y))dy \right|$$

$$\leq 2 \sup_{\substack{\alpha \in B_n(F) \\ \alpha \leq y \leq \alpha_*}} |\alpha_* - \alpha| (F^{-1}(y))^2 + \sup_{\substack{\alpha \in B_n(F) \\ \alpha \leq y \leq 1-\alpha}} |F_n^{-1}(1-y)| |F_n^{-1}(y) - F^{-1}(y)|$$

$$+ \sup_{\substack{\alpha \in B_n(F) \\ \alpha \le y \le 1-\alpha}} |F^{-1}(y)||F_n^{-1}(1-y) - F^{-1}(1-y)|$$

which converges to zero using (3.3) and Lemma 6.2. This completes the proof of Theorem 4.1.

Acknowledgement. I would like to thank P.K. Bhattacharya for suggesting this problem to me.

References

Beran, R. (1974). Asymptotically efficient adaptive rank estimates in location models. *Ann. Statist.* 2 63-74.

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* 71 352-360.

Bowman, A.W, Hall, P., and Titterington, D.M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika* 71 341-351.

Chow, Y.-S., Geman, S., and Wu, L.-D. (1983). Consistent cross-validated density estimation. *Ann. Statist.* 11 25-38.

Csörgö, M. (1983). *Quantile Processes with Statistical Applications*. SIAM, Philadelphia.

Dvoretsky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function of the multinomial estimator. *Ann. Math. Statist.* 27 642-669.

Eastment, H.T., and Krzanowski, W.J. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, 24 73-77.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* 70 320-328.

Hall, P. (1981). On nonparametric multivariate binary discrimination. *Biometrika* 68 287-294.

Hogg, R.V. (1974). Adaptive robust procedures. *J.Amer.Statist.Assoc.* 69 909-927.

Jaeckel, L.A. (1971). Some flexible estimates of location. *Ann.Math. Statist.* 42 1540-1552.

Li, K.-C. (1987). Asymptotic optimality for C_P,C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann.Statist.* 15 958-975.

Parzen, E. (1979). Nonparametric statistical data modelling. *J.Amer. Statist.Assoc.* 74 105-131.

Prescott, P. (1978). Selection of trimming proportions for robust adaptive trimmed means. *J.Amer.Statist.Assoc.* 73 133-140.

Sacks, J. (1975). An asymptotically efficient sequence of estimators of a location parameter. *Ann.Statist.* 3 285-298.

Stone, C.J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann.Statist.* 3 267-284.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J.Roy.Statist. Soc. Ser B.* 36 111-147.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika* 64 29-35.